# Robust estimation of microbial diversity in theory and in practice

Bart Haegeman[1], Jérôme Hamelin[2], John Moriarty[3], Peter Neal[4], Jonathan Dushoff[5], Joshua S. Weitz[6]

[1]*Centre for Biodiversity Theory and Modelling, Experimental Ecology Station, Centre National de Recherche Scientifique, Moulis, France;* [2]*INRA, UR50, Laboratoire de Biotechnologie de l'Environnement, Narbonne, France;* [3]*School of Mathematics, University of Manchester, Manchester, United Kingdom;* [4]*Department of Mathematics and Statistics, University of Lancaster, Lancaster, United Kingdom;* [5]*Department of Biology and Institute of Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada;* [6]*School of Biology and School of Physics, Georgia Institute of Technology, Atlanta, Georgia, United States of America*

**Quantifying diversity is of central importance for the study of structure, function and evolution of microbial communities. The estimation of microbial diversity has received renewed attention with the advent of large-scale metagenomic studies. Here, we consider what the diversity observed in a sample tells us about the diversity of the community being sampled. First, we argue that one cannot reliably estimate the absolute and relative number of microbial species present in a community without making unsupported assumptions about species abundance distributions. The reason for this is that sample data do not contain information about the number of rare species in the tail of species abundance distributions. We illustrate the difficulty in comparing species richness estimates by applying Chao's estimator of species richness to a set of *in silico* communities: they are ranked incorrectly in the presence of large numbers of rare species. Next, we extend our analysis to a general family of diversity metrics ("Hill diversities"), and construct lower and upper estimates of diversity values consistent with the sample data. The theory generalizes Chao's estimator, which we retrieve as the lower estimate of species richness. We show that Shannon and Simpson diversity can be robustly estimated for the *in silico* communities. We analyze nine metagenomic data sets from a wide range of environments, and show that our findings are relevant for empirically-sampled communities. Hence, we recommend the use of Shannon and Simpson diversity rather than species richness in efforts to quantify and compare microbial diversity.**

Accepted paper in press at *The ISME Journal*; doi:10.1038/ismej.2013.10

**Subject category:** Microbial population and community ecology

**Keywords:** Chao estimator; Hill diversities; metagenomics; Shannon diversity; Simpson diversity; species abundance distribution

# Introduction

Species diversity is a crucial property of ecological communities: it is the primary descriptor of community structure, and it is generally believed to be a major determinant of the functioning and the dynamics of ecological communities (Wilson, 1999; Loreau *et al.*, 2001; Ives and Carpenter, 2007; Loreau, 2010). Therefore, diversity measurement is often a first step in characterizing an ecological community (Brose *et al.*, 2003; Magurran, 2004; Gotelli and Colwell, 2011). Because an exhaustive census of the community is usually not feasible, community diversity must be inferred from the diversity observed in a sample taken from the community. The inference problem can be difficult, especially when community diversity is believed to be very large (Engen, 1978; Bunge and Fitzpatrick, 1993; Mao and Colwell, 2005).

Diversity measurement is particularly challenging for microbial communities (Hughes *et al.*, 2001; Bohannan and Hughes, 2003; Kemp and Aller, 2004; Schloss and Handelsman, 2005; Sloan *et al.*, 2008; Bunge, 2009; Øvreås and Curtis, 2011). First, it should be recalled that there is no unambiguous way to define microbial "species" (Stackebrandt *et al.*, 2002). Here we use the term species pragmatically to mean an operationally determined taxonomic unit (e.g., 97% identity of 16S rRNA (Schloss and Handelsman, 2005)). However measured, the species diversity of microbial communities is usually much larger than that of communities of larger organisms. Moreover, the number of organisms in microbial communities is typically many orders of magnitude larger than the number of organisms in plant or animal communities (Whitman *et al.*, 1998). This leads to severe sampling problems. Although metagenomic approaches allow for impressively large sample size (Huber *et al.*, 2007; Roesch *et al.*, 2007; Rusch *et al.*, 2007), even these huge samples correspond to a tiny fraction of the community being sampled. Hence, for microbial community samples, community diversity is generally much larger than sample diversity. This disparity between community and sample leads to a challenge that we address here: how can microbial diversity be estimated robustly?

One popular approach to circumvent the sampling problem is to assume that the species abundance distribution of the community belongs to a specific family (for example, the family of lognormal distributions) (Curtis *et al.*, 2002; Hong *et al.*, 2006; Schloss and Handelsman, 2006; Quince *et al.*, 2008). Such an assumption fills in the information about the community missing in the data and leads to precise diversity estimates. But the validity of the estimates depends crucially on the choice of the species abundance distribution family. This choice cannot be verified empirically because the sample data do not contain sufficient information about the community structure. In fact, many distribution families yield extrapolated community structures that are consistent with the sample data. Here we show that the extrapolation approach has intrinsic limitations.

Other methods for diversity estimation have been proposed. For example, proposals have been made to extrapolate the rarefaction curve beyond the actual sample size (Gotelli and Colwell, 2001; Colwell *et al.*, 2004), or to assume a particular distribution for the community diversity over taxonomic levels (May, 1988; Mora *et al.*, 2011). Eventually, also these methods are limited by the lack of information about the community structure in the sample data. Rather than filling this gap by unverifiable assumptions, here we ask what can (and cannot) be inferred from the sample data alone. An interesting step in this direction is given by the popular Chao estimator (Chao, 1984; Shen *et al.*, 2003; Chao *et al.*, 2009). Chao's estimate can be interpreted as a *lower estimate* of the species richness consistent with the data. We take the estimation strategy underlying Chao's estimator a step further, and construct lower and upper estimates for a general family of community diversities, including species richness, Shannon diversity and Simpson diversity (Hill, 1973). The unification we propose here represents a robust approach to estimating microbial diversity in theory and in practice.

# Materials and methods

### Data sets

The data sets used in this paper were downloaded from the supplementary material of Quince *et al.* (2008). The abundance data used in Figure 1 correspond to 16S rDNA sequences obtained from a bacterial soil community (sample "Brazil" in Roesch *et al.* (2007)). The abundance data used in Figure 5 correspond to 16S rDNA sequences obtained from a bacterial seawater community from the upper ocean (Rusch *et al.*, 2007), from four bacterial soil communities (Roesch *et al.*, 2007), and from bacterial and archaeal seawater communities from two hydrothermal vents (Huber *et al.*, 2007).

### Rank-abundance curves

We represent the species abundance distribution of a community as a rank-abundance curve, that is, we arrange the species in decreasing order of community abundance, and plot species abundance as a function of species rank. We use logarithmic scales for both axes of the rank-abundance curves, so that a community with power-law abundance distribution is represented as a straight line (the slope is equal to the power-law exponent), see Figure 2A. We constructed the communities of Figure 1 by using a piecewise linear parametrization of the rank-abundance curve. Hence, the species abundance distributions consist of power-law segments with different exponents.

### Rarefaction curves

We define $S_m$ as the expected number of species in a sample of $m$ individuals taken from the community (sampling with replacement). The rarefaction curve of the community is the plot of the number of species $S_m$ as a function of the sample size $m$. It is important to distinguish the community rarefaction curve from the rarefaction curve estimated from sample data. For a sample of size $M$ taken from the community, the part of the rarefaction curve corresponding to $S_m$ with $m \leq M$ can be estimated by subsampling the sample data. The same approach fails for the part of the rarefaction curve corresponding to $S_m$ with $m > M$. In that case the rarefaction curve has to be extrapolated, introducing large estimation uncertainty. We studied two extreme extrapolation scenarios: one for the slowest (i.e., smallest slope) and one for the fastest (i.e., largest slope) increase of the rarefaction curve compatible with the sample data, see Figure 3.

### Hill diversities

The Hill diversities, defined in Equation (3), can be computed if the community abundances are known. If only sample data are available, Hill diversities have to be estimated. We consider sampling with replacement, and denote by $M$ the sample size and by $F_k$ the number of species sampled $k$ times. We developed an estimation procedure that exploits the link between Hill diversities $D_\alpha$ and the rarefaction curve $S_m$. The lower estimate $\widehat{S}_m^-$ of the rarefaction curve,

$$\widehat{S}_m^- = \begin{cases} \sum_{k\geq 1} F_k \Big(1 - \frac{\binom{M-k}{m}}{\binom{M}{m}}\Big) & \text{if } m \leq M \\ S_{\text{obs}} + \frac{F_1^2}{2F_2}\Big(1 - \big(1 - \frac{2F_2}{MF_1}\big)^{m-M}\Big) & \text{if } m > M, \end{cases}$$

yields the lower estimate of the Hill diversity,

$$\widehat{D}_\alpha^- = \Big( \sum_{m=1}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(1-\alpha)}\,\widehat{S}_m^- \Big)^{\frac{1}{1-\alpha}}, \tag{1}$$

where $\Gamma$ denotes the gamma function. Similarly, the upper estimate of the rarefaction curve,

$$\widehat{S}_m^+ = \begin{cases} \sum_{k\geq 1} F_k \Big(1 - \frac{\binom{M-k}{m}}{\binom{M}{m}}\Big) & \text{if } m \leq M \\ S_{\text{obs}} + \frac{NF_1}{M}\Big(1 - \big(1-\frac{1}{N}\big)^{m-M}\Big) & \text{if } m > M, \end{cases}$$

with $N$ the (estimated) community size, yields the upper estimate of the Hill diversity,

$$\widehat{D}_\alpha^+ = \Big( \sum_{m=1}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(1-\alpha)}\,\widehat{S}_m^+ \Big)^{\frac{1}{1-\alpha}}. \tag{2}$$

The estimators (1) and (2) can be computed with the Matlab code in the Supplementary Information, and were used to generate Figures 4 and 5.

# Results

### Species richness cannot be estimated from sample data alone

We are interested in estimating the diversity of a community based on the composition of a sample taken from the community. Our approach is to reconstruct community structures, i.e., species abundance distributions, from the sample data. For the example data set of Figure 1, we find that a wide range of communities are consistent with the sample data. The reconstructed communities have vastly different numbers of species, differing by two orders of magnitude, implying that estimating species richness is subject to large biases.

We claim that sample data is always consistent with very different community structures. To establish this claim we study the link between the rare species tail of the community and the sample data, summarized by the rarefaction curve. A computation in Supplementary Text S1 shows that the rarefaction curve up to sample size $M$ is insensitive to the abundance distribution of species with relative abundance well below $\frac{1}{M}$. For concreteness we set a relative abundance threshold at $\frac{1}{50\,M}$, and we call the species with larger and smaller relative abundance than this threshold the "non-rare" and "rare" species, respectively. The computation shows that the rarefaction curves does not depend on the abundance distribution of the rare species. Changes in the rare species tail, such as increasing the number of rare species by several orders of magnitude (but keeping the total abundance of rare species constant), does not affect the sample data. As a consequence, estimating species richness is intrinsically problematic.

Note that we use a statistical definition of rarity which depends on the sampling effort $M$; the set of rare species gets smaller when sampling gets deeper. This contrasts with the ecological concept of rarity, a community property independent of sample size (Pedrós-Alió, 2006; Sogin *et al.*, 2006), see the Discussion section.

To further illustrate the theoretical result we reconsider the reconstructed communities of Figure 1. The communities have the same abundance distribution of the non-rare species. In each community the set of rare species occupies 0.5% of the total community abundance, explaining why the corresponding rarefaction curves coincide, see Figure 1D. Nevertheless, the number of rare species differs by two orders of magnitude. Another example of *in silico*

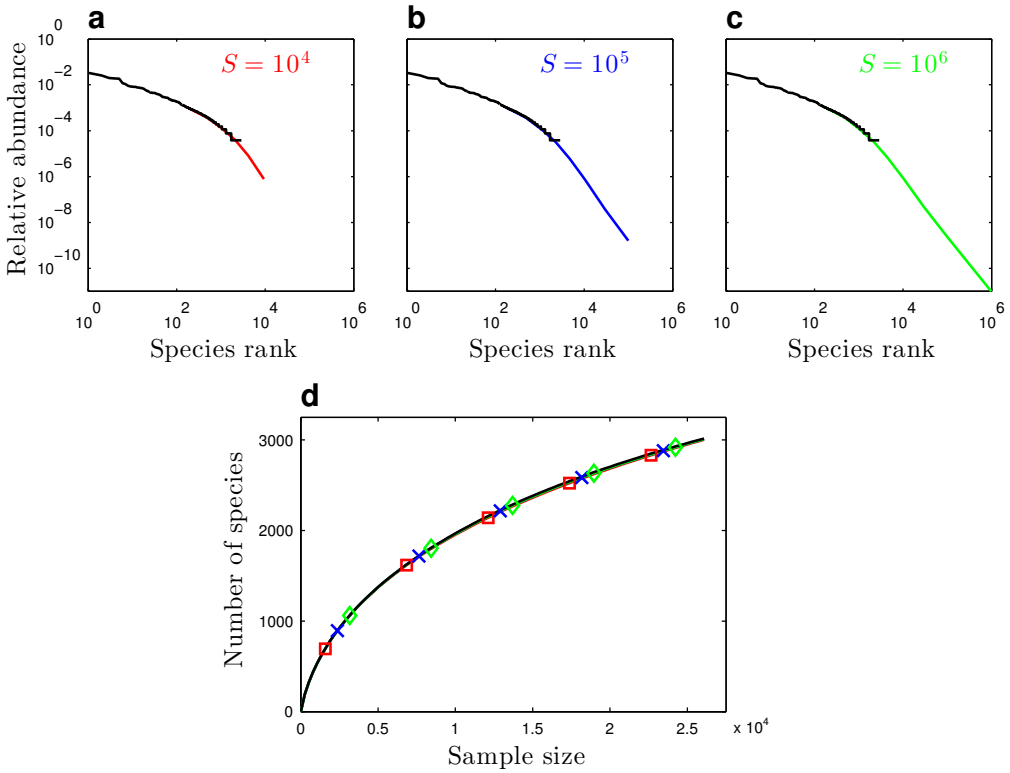


Figure 1: Empirical sample data are consistent with very different communities. We consider the abundance data of a sample taken from a bacterial soil community (sample "Brazil" in Roesch *et al.* (2007)). The sample consists of 26079 individuals belonging to 2880 species. We tried to reconstruct the community from which the sample was taken. Panels a–c show the rank-abundance curve of three such reconstructed communities. The first community (panel a, in red) has $10^4$ species; the second community (panel b, in blue) has $10^5$ species; the third community (panel c, in green) has $10^6$ species. For each of the three reconstructions the community rank-abundance curve is an extension of the sample rank-abundance curve (in black). We claim that each of the three reconstructed communities is compatible with the sample data. This can be seen from the rarefaction curves in panel d: the rarefaction curve for the sample data (black line) coincides with the rarefaction curves for the reconstructed communities (red line with squares for community in panel a, blue line with $\times$-marks for community in panel b, and green line with diamonds for community in panel c). Because the sample data are consistent with very different values of the community richness, the community richness cannot be estimated from the sample data.

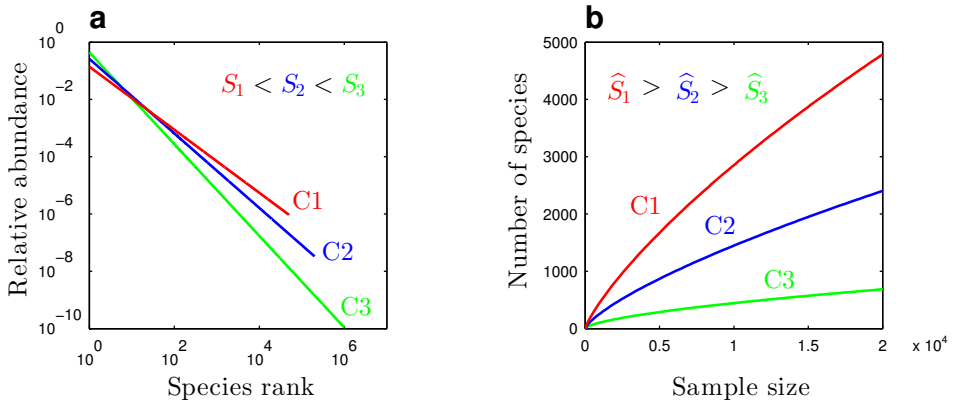


Figure 2: Estimated species richness does not rank correctly communities. We generated three community abundance distributions, the rank-abundance curves of which are shown in panel a. Community C1 (red) has the smallest number of species; community C3 (green) has the largest number of species. The rarefaction curves of the three communities up to sample size 2 $10^4$ are shown in panel b. Based on the rarefaction data, one would conclude that community C1 is the most diverse and community C3 the least diverse. Hence, the ranking of the communities according to their observed diversity is inverted compared to the ranking according to their true diversity. This observation is confirmed when applying Chao's estimator to sample data. Community C1 is estimated to have 10 times more species than community C3, whereas in reality community C1 has 20 times less species than community C3. See Supplementary Table S1 for the numerical data of the communities.

communities with very different rare species tails but with the same rarefaction curve is shown in Supplementary Figure S1.

We conclude that sample data do not allow us to distinguish communities with very different rare species tails. The insensitivity of the rarefaction curve to rare species implies that it is difficult or impossible to reliably estimate the community species richness from sample data alone.

### Relative species richness cannot be estimated from sample data alone

We have shown that the number of species in a community cannot be reliably estimated from sample data. A related question is whether sample data can be used to rank different communities according to their number of species. In this section we show that this cannot be done without additional assumptions.

We present an explicit example to illustrate the use of sample data to rank communities, see Figure 2. We consider three communities which differ widely in species richness: community C1 has 20 times fewer species than community C3. We construct the initial arcs of these rarefaction curves, see Figure 2B. Surprisingly, the rarefaction curves suggest that community C1 is the most diverse, and community C3 the least diverse. We therefore expect that any estimator of species richness ranks the communities in the inverse order of their true species richness. Indeed, Chao's estimator predicts that community C1 has almost 10 times as many species as community C3 (see Supplementary Table S1; values are averaged over sample randomness).

To understand the incorrect ranking we take a closer look at the communities in Figure 2A. We explained, in the previous section, that sample data are insensitive to rare species. When

we compare the number of non-rare species in the communities (species with relative abundance above $10^{-6}$), we find that community C1 has 15 times more non-rare species than community C3. This explains why the sample data suggest that community C1 is the most diverse. Community C1 has a large number of non-rare species combined with a relatively small number of rare species. In contrast, community C3 has a relatively small number of non-rare species combined with a very large number of rare species. This explains the discrepancy between true number of species, mainly determined by the rare species, and estimated number of species, determined by the non-rare species.

The example of Figure 2 indicates a general problem: relative species richness cannot be reliably estimated. The problem is due to the same mechanism as the one identified in the previous section. Sample data cannot be used to rank communities according to their number of species because sample data do not contain information about the number of rare species.

**Some generalized diversities can be estimated from sample data alone**

Altough insensitive to rare species, sample data do contain information about the community structure. In this section we demonstrate that diversity indices that are weakly dependent on rare species can be estimated from sample data.

Diversity is a broader notion than species richness. Alternative definitions of diversity have been proposed in which rare species contribute less than common species. These alternative diversities account not only for species richness but also for the evenness of the community structure. Examples are the Shannon diversity index (Shannon, 1948) and the Simpson diversity index (Simpson, 1949). Here we study a family of generalized diversities, the Hill diversities $D_\alpha$ (Hill, 1973) that includes these two examples as well as species richness as special cases. For a community consisting of $S$ species with relative abundances $p_1, p_2, \ldots, p_S$, the Hill diversities are defined by

$$D_\alpha = \left( \sum_{i=1}^{S} p_i^\alpha \right)^{\frac{1}{1-\alpha}}. \tag{3}$$

We obtain a Hill diversity for each value of the parameter $\alpha$. For $\alpha = 0$ the species are weighted equally in the sum of Equation (3) (each term is equal to one), and $D_0 = S$, i.e., $D_0$ is equal to species richness. For $\alpha > 0$ the species are not weighted equally. Instead, a rare species contributes less than a common species. For larger values of $\alpha$ the weighting is more unequal, see Supplementary Text S2. As an extreme case, only the most abundant species contributes in the limit $\alpha \to \infty$. The Hill diversity of order 1 is related to the Shannon diversity index (note that Definition (3) should be understood as $D_1 = \lim_{\alpha \to 1} D_\alpha$) and the Hill diversity of order 2 is related to the Simpson concentration index. The Hill diversity for a community in which all $S$ species have the same relative abundance $p_i = \frac{1}{S}$ is equal to $D_\alpha = S$ for any value of the parameter $\alpha$. This indicates that any Hill diversity $D_\alpha$ can be considered as an effective number of species (Hill, 1973; Jost, 2006), which facilitates the interpretation of estimated diversity values and allows us to compare the estimation properties of different Hill diversities.

As $\alpha$ increases the Hill diversities are increasingly insensitive to the tail of rare species and are more strongly determined by the non-rare species, see Supplementary Figure S2. Hence, we expect that they are more accurately estimated from sample data. A mathematical link between the Hill diversities and the rarefaction curve further indicates which Hill diversities can be estimated from sample data. In Supplementary Text S3 we show that any Hill diversity $D_\alpha$ can be expressed in terms of the rarefaction curve. The Hill diversity $D_2$ is related to the initial slope of the rarefaction curve (Lande *et al.*, 2000). Thus, for $\alpha$ close to 2, the Hill diversity $D_\alpha$ depends on the part of the rarefaction curve for small sample size. For smaller

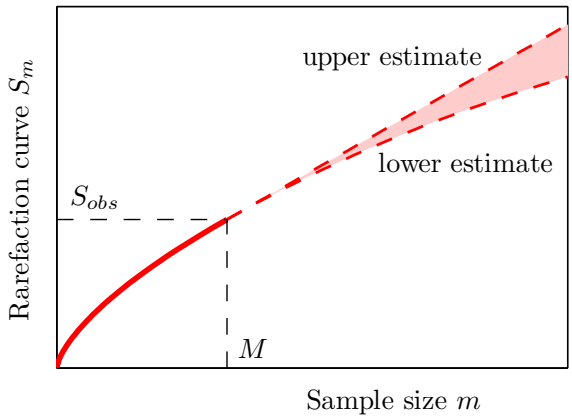


Figure 3: Extrapolating the rarefaction curve. The Hill diversity estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ are based on reconstructions of the rarefaction curve $S_m$ from sample data. For a sample of size $M$, the rarefaction curve $S_m$ for $m \leq M$ can be estimated by subsampling (red full line). If the sample size $M$ is large, the estimator has small uncertainty. The rarefaction curve $S_m$ for $m > M$ can be estimated by extrapolating the sample data beyond the sample size $M$. Different extrapolation scenarios are compatible with the sample data. We consider two extreme scenarios (red dashed lines). A lower estimate is obtained by assuming that unobserved species are approximately as rare as the rarest observed species. An upper estimate is obtained by assuming that unobserved species are represented in the community by one individual. The difference between the two extremes quantifies the uncertainty of the extrapolation, shown as the red shaded region. The uncertainty increases rapidly for $m \gg M$.

$\alpha$, the Hill diversity $D_\alpha$ depends on the rarefaction curve for increasingly large sample size. The Hill diversity $D_0$ is equal to species richness, which can be obtained as the limit of the rarefaction curve for infinite sample size.

These observations have important implications for the diversity estimation problem. We suppose that sample data of size $M$ are given, and we try to estimate the rarefaction curve at sample size $m$. The community rarefaction curve for sample sizes $m \leq M$ can be estimated in an unbiased manner by subsampling the sample data, but for $m > M$ the rarefaction curve can only be estimated based on extrapolation. This leads to increasingly biased estimates as $m$ increases. Hence, we reach the following conclusions. On one hand, Hill diversities that depend on the initial part of the rarefaction curve, that is, $D_\alpha$ for $\alpha$ close to 2, can be estimated robustly. On the other hand, Hill diversities that depend on the part of the rarefaction curve for large sample size, that is, $D_\alpha$ for $\alpha$ close to 0, cannot be estimated robustly. We now seek to make this classification of community diversities more precise.

### Estimators for Hill diversities

We have argued that the Hill diversities $D_\alpha$ with $\alpha$ close to 2 can be estimated accurately, and that the Hill diversities $D_\alpha$ with $\alpha$ close to 0 cannot be estimated accurately. In this section we introduce and study estimators for the set of Hill diversities $D_\alpha$ with $0 \leq \alpha \leq 2$.

We have shown that a wide variety of communities may be consistent with any given sample data. Here we look for two extreme members of this set of reconstructed communities. We construct a lower estimate of the diversity, $\widehat{D}_\alpha^-$, by assuming that unobserved species are approximately as rare as the rarest observed species. We construct an upper estimate of the

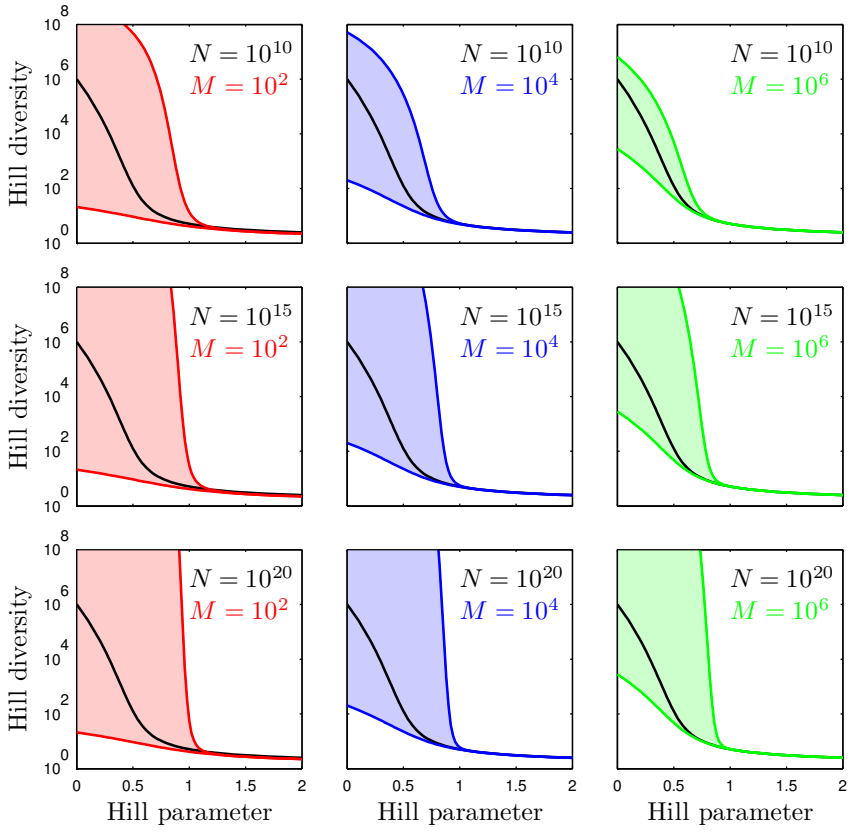


Figure 4: Estimated Hill diversities for *in silico* communities. We generated samples from a community with power-law abundance distribution ($S = 10^6$, $z = 2$) and evaluated the estimators $\widehat{D}_\alpha^+$ and $\widehat{D}_\alpha^-$ for the Hill diversity $D_\alpha$. We consider three sample sizes $M$ (in columns: $M = 10^2, 10^4, 10^6$) and three community sizes $N$ (in rows: $N = 10^{10}, 10^{15}, 10^{20}$). The shaded range between $\widehat{D}_\alpha^+$ and $\widehat{D}_\alpha^-$ indicates the estimation uncertainty. The true Hill diversity $D_\alpha$ of the community is plotted in black. The Hill diversities between $\alpha = 1$ (Shannon) and $\alpha = 2$ (Simpson) are correctly estimated even for small sample size $M$. The estimates of Hill diversities less than $\alpha = 1$, including $\alpha = 0$ (species richness), are characterized by large uncertainty.

diversity, $\widehat{D}_\alpha^+$, by assuming that unobserved species are represented in the community by a single individual. We first extrapolate the rarefaction curve based on these assumptions, see Figure 3, and then use the extrapolated curves to calculate the Hill diversities. The detailed construction of the estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ is presented in Supplementary Texts S3, S4 and S5. A summary of the estimator formulas can be found in the Materials and Methods section. We provide Matlab code to compute the estimators in the Supplementary Information.

Two properties follow directly from the definition of the estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$, see Supplementary Text S5. First, the lower estimate $\widehat{D}_0^-$ for species richness is equal to Chao's estimator. Hence, the lower estimate $\widehat{D}_\alpha^-$ generalizes Chao's estimator for Hill diversities $D_\alpha$ with $\alpha > 0$. Second, the estimators for Simpson diversity $D_2$ coincide, $\widehat{D}_2^- = \widehat{D}_2^+$. This corresponds to the existence of an unbiased, non-parametric estimator for the Simpson concentration index, and confirms that Simpson diversity $D_2$ is particularly easy to estimate, even for small sample size $M$. Note that the lower estimate can be computed from the sample data alone, but the upper estimate also requires an estimate of the community size $N$.

In Figure 4 we apply the estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ to sample data from an *in silico* community. For $\alpha > 1$ the lower and upper estimates almost coincide, so that the Hill diversities $D_\alpha$ with $\alpha > 1$, and in particular Simpson diversity $D_2$, may be estimated with small error. This holds for any sample size $M$ (as small as $M = 100$) and any community size $N$. For $\alpha < 1$ the upper estimate increases steeply, so that the estimation uncertainty of the Hill diversities $D_\alpha$ with $\alpha$ small, and in particular species richness $D_0$, is very large. This holds for any sample size $M$ (as large as $M = 10^6$) and any community size $N$ much greater than $M$. The effect of sample size $M$ and community size $N$ is only pertinent for $\alpha$ close to 1. For these values of $\alpha$ the range between the lower and upper estimates narrows with increasing sample size $M$ and decreasing community size $N$, so that increasingly accurate estimates are obtained for Shannon diversity $D_1$.

We observe the same behavior when applying the Hill diversity estimators to empirical sample data, see Figure 5. We applied the estimators to nine metagenomic data sets from a wide range of environments: soil samples at four locations (Roesch *et al.*, 2007), a seawater sample from the upper ocean (Rusch *et al.*, 2007) and seawater samples at two deep-sea vent locations (Huber *et al.*, 2007). The estimators exhibit the same patterns as for the *in silico* community studied in Figure 4. The Hill diversities $D_\alpha$ for $\alpha \geq 1$, including Shannon and Simpson diversity, can be estimated reliably. For small $\alpha$ the estimation uncertainty is very large, that is, Hill diversities close to species richness cannot be estimated reliably. The dependence of the estimation accuracy on the (estimated) community size $N$ is weak, see Supplementary Figure S4. These observations show that our analysis for *in silico* communities is relevant for real communities as well.

# Discussion

We have argued that the estimation of species richness is intrinsically problematic. We have provided evidence in three different but related ways. First, we have shown that it is possible to add a large number of rare species to the community without significantly affecting its statistical properties under fixed-size sampling, see Figure 1. As the number of added rare species can be large, the estimation uncertainty of the number of species is large as well. Second, we have discussed an exact relationship between the community rarefaction curve and the set of Hill diversities. Hill diversities close to Simpson's are based on the initial part of the rarefaction curve, which can be reliably interpolated from sample data. Hill diversities beyond Shannon's, and species richness in particular, depend on parts of the rarefaction curve orders of magnitude beyond the actual sample size, whose estimation requires unverifiable extrapolation. Third, we have constructed two estimators related to the Hill diversities, delimiting the range in which each true Hill diversity is expected to lie. This range is relatively narrow for diversities from Simpson's to Shannon's, but it diverges for diversities towards species richness, see Figures 4 and 5. Hence, the estimation uncertainty of species richness is intrinsically large.

We have also studied a weaker form of species richness estimation, namely, whether communities can be ranked according to species richness based on sample data. We have argued that also in this case the sample data are not sufficiently informative. The example shown in Figure 2 is interesting, because the community ranking based on estimated species richness, although completely different from the ranking based on true richness, is the same as the ranking based on true Simpson or Shannon diversity, see Supplementary Table S1. This observation can be understood intuitively. The insensitivity of the species richness estimator to the very rare species in the community is shared by the Simpson and Shannon diversity, but not by the community species richness. In fact, different diversity estimators often yield the same community ranking (Shaw *et al.*, 2008). This should not be interpreted as an indication of the validity of the ranking for species richness; the ranking based on true species richness

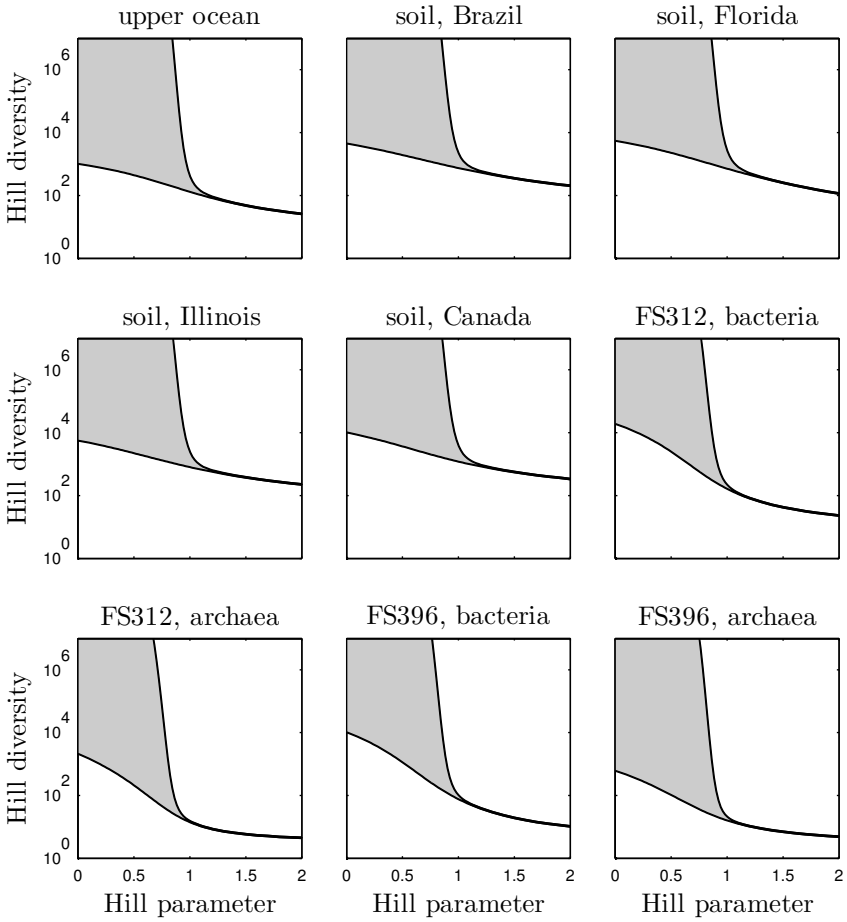


Figure 5: Estimated Hill diversities for natural microbial communities. We observe the same behavior as for the *in silico* generated data sets of Figure 4: for $\alpha \geq 1$ the Hill diversity $D_\alpha$ can be estimated accurately; for $\alpha < 1$ the estimation of the Hill diversity $D_\alpha$ has large uncertainty. We used the same data sets as Quince *et al.* (2008): a seawater bacterial sample from the upper ocean (Rusch *et al.*, 2007), soil bacterial samples at four locations: Brazil, Florida, Illinois and Canada (Roesch *et al.*, 2007), and seawater samples from deep-sea vents at two locations: FS312 and FS396, separated into bacteria and archaea (Huber *et al.*, 2007). The community size was set to $N = 10^{15}$ for illustration; results are robust to changes in community size (see Supplementary Figure S4).

can be completely different. Communities should only be ranked according to community properties that can be estimated reliably.

The intrinsic problem of species richness estimation can be unlocked by introducing more information in the estimation procedure. Obviously, the reliability of the estimate crucially depends on the reliability of the additional information. For example, assuming a family of abundance distributions (for example, lognormal) can lead to species richness estimates with small uncertainty (Schloss and Handelsman, 2005; Hong *et al.*, 2006; Quince *et al.*, 2008). But both the estimate and the uncertainty are conditional on the assumed distribution family. In particular, assuming a species abundance distribution also fixes the rare species tail and, as we have argued, the sample data contain little information about the rare species tail. Hence, the choice of distribution family is arbitrary. Still, this choice strongly affects the species richness estimate. We believe this to be a serious problem for this approach to diversity estimation.

Other assumptions have been introduced to make diversity estimation manageable. Some regularity has been observed in the distribution of diversity over coarse taxonomic groups (Mora *et al.*, 2011). This regularity can be assumed down to the species level to guide the estimation of species richness. Clearly, the approach depends crucially on the unverifiable validity of the extrapolation. More generally, this and other approaches attempt to reduce the wide range of diversity values consistent with the data to a single value. This implies that the reduction step is based on detailed information not contained in the sample data. Such an approach is necessarily very sensitive to the detailed assumptions, and therefore not robust.

Mao and Colwell (2005) pointed out that rare species pose a serious problem for estimating species richness. In this paper we have shown a practical way forward by quantifying the range of diversity values consistent with the data. The latter idea underlies our construction of lower and upper estimates of community diversity, and is also crucial for Chao's estimator of species richness (Chao, 1984). This estimator does not attempt to directly assess true species richness, but rather approximates the lowest species richness consistent with the sample data. In many practical cases this indirect estimation is the most informative claim that can be made about species richness.

Different studies have highlighted the role of rare species in microbial communities (Dykhuizen, 1998; Pedrós-Alió, 2006; Sogin *et al.*, 2006; Pedrós-Alió, 2007; Huber *et al.*, 2007; Gobet *et al.*, 2010). We have argued that sample data contain limited information about the rare species tail of the community. For example, the total number of rare species cannot be estimated. However, an estimator for the relative abundance of unobserved species is available, see Supplementary Text S4. For the data sets we have analyzed the estimated relative abundance ranges from 0.1% to 5%, see Supplementary Table S2. These estimates depend on sample size. It might be more practical to use a notion of rarity that is independent of sample size. For example, we could call a species rare if its community abundance is below a certain threshold value (for example, relative abundance below $10^{-4}$). We plan to address the problem of estimating the relative abundance of rare species in a sample-independent fashion as part of future work.

In this paper we have only considered taxonomic diversity. Other notions of diversity such as functional and phylogenetic diversity are becoming increasingly popular (Horner-Devine and Bohannan, 2006; Lozupone and Knight, 2007; Green *et al.*, 2008). Our study suggests that any diversity metrics that strongly depend on rare species will be difficult or impossible to estimate robustly. It is interesting to note that other measurement techniques for microbial diversity are confronted with limitations similar to those of the sample-based techniques discussed in this paper. The reassociation kinetics of community DNA are affected by community diversity (Torsvik *et al.*, 1990; Gans *et al.*, 2005), but it has been argued that not species richness, but Simpson and Shannon diversity can be estimated from the data (Haege-

man *et al.*, 2008). Fingerprinting techniques provide snapshots of the community structure (Fromin *et al.*, 2002): in this context also, the estimation of species richness seems to be impossible for highly diverse communities (Loisel *et al.*, 2006; Bent and Forney, 2008), but preliminary results indicate that accurate estimators can be constructed for Simpson diversity. Estimates of the total number of genes in a species, i.e., the pan genome size, has been estimated from a small number of sample genomes (Tettelin *et al.*, 2005), but it is has been argued that these estimates are not robust and that similarity-based metrics should be used instead (Kislyuk *et al.*, 2011).

These findings together with those of this paper make a strong case for the versatility of generalized diversities for the analysis of microbial diversity estimation. They can be interpreted as effective number of species giving greater weight to common species (Hill, 1973; Jost, 2006), and have superior estimation properties compared to species richness. We recommend the use of Shannon and Simpson diversity to quantify and compare microbial taxonomic diversity.

## Acknowledgments

Financial support for B.H. and J.H was provided by the DISCO project from the French National Research Agency (ANR, project number AAP215-SYSCOMM-2009), and for B.H, J.H., J.M. and P.N. by an Alliance grant from the British Council and the French Foreign Affairs Ministry (project number 22732SJ). J.S.W. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

## References

Bent SJ, Forney LJ. (2008). The tragedy of the uncommon: understanding limitations in the analysis of microbial diversity. *ISME J* **2**: 689–695.

Bohannan BJM, Hughes JB. (2003). New approaches to analyzing microbial biodiversity data. *Curr Opin Microbiol* **6**: 282–287.

Brose U, Martinez ND, Williams RJ. (2003). Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology* **84**: 2364–2377.

Bunge J. (2009). Statistical estimation of uncultivated microbial diversity. In: Epstein SS (ed). *Uncultivated Microorganisms*, Springer-Verlag, pp 1–18.

Bunge J, Fitzpatrick M. (1993). Estimating the number of species: A review. *J Amer Statist Assoc* **88**: 364–373.

Chao A. (1984). Nonparametric estimation of the number of classes in a population. *Scand J Statist* **11**: 265–270.

Chao A, Colwell RK, Lin CW, Gotelli NJ. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* **90**: 1125–1133.

Colwell RK, Mao CX, Chang J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* **85**: 2717–2727.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**: 10494–10499.

Dykhuizen DE. (1998). Santa Rosalia revisited: why are there so many species of bacteria? *Antonie Van Leeuwenhoek* **73**: 25–33.

Engen S. (1978). *Stochastic Abundance Models*. Chapman & Hall.

Fromin N, Hamelin J, Tarnawski S, Roesti D, Jourdain-Miserez K, Forestier N, *et al.* (2002). Statistical analysis of denaturing gel electrophoresis (DGE) fingerprinting patterns. *Environ Microbiol* **4**: 634–643.

Gans J, Wolinsky M, Dunbar J. (2005). Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* **309**: 1387–1390.

Gobet A, Quince C, Ramette A. (2010). Multivariate cutoff level analysis (MultiCoLA) of large community data sets. *Nucl Acids Res* **38**: e155.

Gotelli NJ, Colwell RK. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* **4**: 379–391.

Gotelli NJ, Colwell RK. (2011). Estimating species richness. In: Magurran AE, McGill BJ (eds). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, pp 39–54.

Green JL, Bohannan BJM, Whitaker RJ. (2008). Microbial biogeography: From taxonomy to traits. *Science* **320**: 1039–1043.

Haegeman B, Vanpeteghem D, Godon JJ, Hamelin J. (2008). DNA reassociation kinetics and diversity indices: richness is not rich enough. *Oikos* **117**: 177–181.

Hill MO. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54**: 427–432.

Hong SH, Bunge J, Jeon SO, Epstein SS. (2006). Predicting microbial species richness. *Proc Natl Acad Sci USA* **103**: 117–122.

Horner-Devine MC, Bohannan BJM. (2006). Phylogenetic clustering and overdispersion in bacterial communities. *Ecology* **87**: S100–S108.

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.

Ives A, Carpenter S. (2007). Stability and diversity of ecosystems. *Science* **317**: 58–68.

Jost L. (2006). Entropy and diversity. *Oikos* **113**: 363–375.

Kemp P, Aller J. (2004). Bacterial diversity in aquatic and other environments: what 16S rDNA libraries can tell us. *FEMS Microbiol Ecol* **47**: 161–171.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. (2011). Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**: 32.

Lande R, DeVries PJ, Walla TR. (2000). When species accumulation curves intersect: implications for ranking diversity using small samples. *Oikos* **89**: 601–605.

Loisel P, Harmand J, Zemb O, Latrille E, Lobry C, Delgenès JP, *et al.* (2006). Denaturing gradient electrophoresis (DGE) and single-strand conformation polymorphism (SSCP) molecular fingerprintings revisited by simulation and used as a tool to measure microbial diversity. *Environ Microbiol* **8**: 720–731.

Loreau M. (2010). *From Populations to Ecosystems: Theoretical Foundations for a New Ecological Synthesis*. Princeton University Press.

Loreau M, Naeem S, Inchausti P, Bengtsson J, Grime JP, Hector A, *et al.* (2001). Biodiversity and ecosystem functioning: Current knowledge and future challenges. *Science* **294**: 804–808.

Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.

Magurran AE. (2004). *Measuring Biological Diversity*. Blackwell Publishing.

Mao CX, Colwell RK. (2005). Estimation of species richness: Mixture models, the role of rare species, and inferential challenges. *Ecology* **86**: 1143–1153.

May RM. (1988). How many species are there on earth? *Science* **241**: 1441–1449.

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B. (2011). How many species are there on earth and in the ocean? *PLoS Biol* **9**: e1001127.

Øvreås L, Curtis TP. (2011). Microbial diversity and ecology. In: Magurran AE, McGill BJ (eds). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, pp 221–236.

Pedrós-Alió C. (2006). Marine microbial diversity: can it be determined? *Trends Microbiol* **14**: 257–263.

Pedrós-Alió C. (2007). Dipping into the rare biosphere. *Science* **315**: 192–193.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283—290.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Schloss PD, Handelsman J. (2005). Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microbiol* **71**: 1501–1506.

Schloss PD, Handelsman J. (2006). Toward a census of bacteria in soil. *PLoS Comput Biol* **2**: e92.

Shannon CE. (1948). A mathematical theory of communication. *Bell System Tech J*, **27**: 379–423 and 623–656.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC, Martiny JBH. (2008). It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Shen TJ, Chao A, Lin CF. (2003). Predicting the number of new species in further taxonomic sampling. *Ecology* **84**: 798–804.

Simpson EH. (1949). Measurement of diversity. *Nature* **163**: 688.

Sloan WT, Quince C, Curtis TP. (2008). The uncountables. In: Zengler K (ed). *Accessing Uncultivated Microorganisms: From the Environment to Organisms and Genomes and Back*. ASM Press, pp 35–54.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stackebrandt E, Frederiksen W, Garrity GM, Grimont PAD, Kämpfer P, Maiden MCJ, *et al.* (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**: 1043–1047.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, *et al.* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proc Natl Acad Sci USA* **102**: 13950–13955.

Torsvik V, Salte K, Sorheim R, Goksoyr J. (1990). Comparison of phenotypic diversity and DNA heterogeneity in a population of soil bacteria. *Appl Environ Microbiol* **56**: 776–781.

Whitman WB, Coleman DC, Wiebe WJ (1998). Prokaryotes: The unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.

Wilson EO. (1999). *The Diversity of Life.* W.W. Norton & Company.

Supplementary Information

# Robust estimation of microbial diversity in theory and in practice

B. Haegeman, J. Hamelin, J. Moriarty, P. Neal, J. Dushoff, J. S. Weitz

**Supplementary Text**



**Supplementary Tables**



**Supplementary Figures**



**Computer code**

# Supplementary Text

## Text S1

### Contribution of rare species to rarefaction curve

We define $S_m$ as the expected number of species in a sample of $m$ individuals taken from the community. The rarefaction curve of the community is the plot of the number of species $S_m$ as a function of the sample size $m$. We consider a community consisting of $S$ species with relative abundance $p_1, p_2, \ldots, p_S$. Then the expected number of sampled species $S_m$ is given by

$$S_m = \sum_{i=1}^{S} \Big(1 - (1-p_i)^m\Big). \tag{S1}$$

It is important to distinguish the community rarefaction curve (S1) from the rarefaction curve estimated from sample data. We consider a sample of size $M$ taken from the community. We denote the number of species observed in the sample by $S_{\mathrm{obs}}$, and the number of species with abundance $k$ in the sample by $F_k$. For $m \le M$ the rarefaction curve $S_m$ can be estimated by taking subsamples of size $m$ out of the sample. The average number of species observed in the subsample (averaged over all subsamples of size $m$) is an estimator for $S_m$,

$$\widehat{S}_m = \sum_{k\ge 1} F_k \left(1 - \frac{\binom{M-k}{m}}{\binom{M}{m}}\right), \qquad m \le M. \tag{S2}$$

This estimator is reliable in the sense that it is unbiased (that is, the expected value of $\widehat{S}_m$ is equal to $S_m$). Moreover, there is no other unbiased estimator with smaller variance. For $m > M$ the estimation of the rarefaction curve is necessarily based on extrapolation, leading to less reliable estimates, especially for $m \gg M$.

We define a species to be rare if its relative abundance is much smaller than $\frac{1}{M}$. This means that a rare species is unlikely to be present in the sample (of size $M$). For concreteness we say that

$$\text{species } i \text{ is rare} \quad \text{if} \quad p_i \le \frac{1}{50\,M}. \tag{S3}$$

Note that our definition of rarity depends on the sample size $M$. The choice of a threshold for rarity is arbitrary, though our results are robust to changes in the constant (which in this case has been set to 50) so long as it is much greater than 1.

We consider the rarefaction curve (S1) up to sample size $M$. The contribution of species $i$ can be written as

$$1 - (1-p_i)^m = \sum_{j=1}^{m} \binom{m}{j} p_i^j\,(1-p_i)^{m-j}, \qquad m \le M.$$

The $j$-th term in this sum is the probability that species $i$ is represented $j$ times in a sample of size $m$. For a rare species $i$ we have $p_i \ll \frac{1}{M} \le \frac{1}{m}$, and the first term dominates the other terms. Hence,

$$1 - (1-p_i)^m \approx m\,p_i\,(1-p_i)^{m-1} \approx m\,p_i, \qquad m \le M.$$

Partitioning the set of species into rare and non-rare species, we get

$$\begin{aligned} S_m &\approx \sum_{\substack{i=1 \\ i \text{ non-rare}}}^{S} \left(1-(1-p_i)^m\right) + \sum_{\substack{i=1 \\ i \text{ rare}}}^{S} m\,p_i \\ &= \sum_{\substack{i=1 \\ i \text{ non-rare}}}^{S} \left(1-(1-p_i)^m\right) + m\,p_{\text{rare}}, \qquad m \le M, \end{aligned} \tag{S4}$$

with $p_{\text{rare}}$ the total relative abundance of the set of rare species in the community.

From Equation (S4) it follows that the rarefaction curve does not depend on the abundance distribution of the rare species, but only on the total abundance of the rare species. This follows directly from Definition (S3): it is unlikely that a rare species will be observed twice in a sample of size $m$ (when $m < M$). Therefore, the contribution of the rare species to the sample species richness depends only on their prevalence in the sample which, in turn, depends only on their prevalence in the community. In particular, rarefaction curves obtained for different abundance distributions of the rare species are indistinguishable, see Figure S1.

# Text S2

## Contribution of rare species to Hill diversities

In the main text we have introduced the Hill diversities $D_\alpha$,

$$D_\alpha = \left( \sum_{i=1}^{S} p_i^\alpha \right)^{\frac{1}{1-\alpha}}. \tag{S5}$$

The Hill diversity of order 1 is defined as the limit $D_1 = \lim_{\alpha \to 1} D_\alpha$, and is related to the Shannon diversity index $H$,

$$D_1 = \mathrm{e}^H \qquad \text{with} \qquad H = \sum_{i=1}^{S} -p_i \ln p_i. \tag{S6}$$

The Hill diversity of order 2 is related to the Simpson concentration index $C$,

$$D_2 = \frac{1}{C} \qquad \text{with} \qquad C = \sum_{i=1}^{S} p_i^2.$$

The Hill diversity of order $\infty$ is related to the relative abundance $p_{\max}$ of the most abundant species,

$$D_\infty = \frac{1}{p_{\max}} \qquad \text{with} \qquad p_{\max} = \max\left\{p_1, p_2, \ldots, p_S\right\}.$$

We consider a community in which the rare species occupy a fraction $p_{\text{rare}}$ of the total community abundance. We study the dependence of the Hill diversity on the number of rare species $S_{\text{rare}}$. Assuming that the rare species have equal abundance, we get

$$\begin{aligned} D_\alpha &= \left( \sum_{\substack{i=1 \\ i \text{ non-rare}}}^{S} p_i^\alpha \; + \sum_{\substack{i=1 \\ i \text{ rare}}}^{S} p_i^\alpha \right)^{\frac{1}{1-\alpha}} \\ &= \left( \sum_{\substack{i=1 \\ i \text{ non-rare}}}^{S} p_i^\alpha \; + \; S_{\text{rare}} \left( \frac{p_{\text{rare}}}{S_{\text{rare}}} \right)^\alpha \right)^{\frac{1}{1-\alpha}} \\ &= \left( \sum_{\substack{i=1 \\ i \text{ non-rare}}}^{S} p_i^\alpha \; + \; p_{\text{rare}}^\alpha \, S_{\text{rare}}^{1-\alpha} \right)^{\frac{1}{1-\alpha}}. \end{aligned} \tag{S7}$$

The first term inside the brackets contains the contribution of the non-rare species. The second term inside the brackets, $p_{\text{rare}}^\alpha \, S_{\text{rare}}^{1-\alpha}$, contains the contribution of the rare species. The contribution of the non-rare species is independent of $S_{\text{rare}}$. For $\alpha > 1$ the contribution of the rare species decreases with $S_{\text{rare}}$ and vanishes for $S_{\text{rare}} \to \infty$. Hence, the rare species contribute only weakly to the Hill diversity $D_\alpha$ for $\alpha > 1$. For $\alpha < 1$ the contribution of the rare species increases with $S_{\text{rare}}$ and diverges for $S_{\text{rare}} \to \infty$. Hence, for sufficiently large $S_{\text{rare}}$ the rare species contribution dominates the Hill diversity $D_\alpha$ for $\alpha < 1$. Note that the relative contribution of the rare to the non-rare species has a power-law dependence on $S_{\text{rare}}$ with exponent $1 - \alpha$. For the Hill diversity $D_1$ the relative contribution of the rare to the non-rare species has a logarithmic dependence on $S_{\text{rare}}$, see (S6).

# Text S3

## Hill diversities and rarefaction curve

We follow Mao (2007) to establish a link between the rarefaction curve $S_m$ and the Hill diversities $D_\alpha$. Rewriting the sum $\sum_i p_i^\alpha$, we get

$$\begin{aligned}
\sum_{i=1}^{S} p_i^\alpha &= \sum_{i=1}^{S} \big(1-(1-p_i)\big)^\alpha \\
&= \sum_{i=1}^{S} \sum_{m=0}^{\infty} \frac{(-1)^m \Gamma(\alpha+1)}{m!\,\Gamma(\alpha-m+1)} (1-p_i)^m \\
&= S \sum_{m=0}^{\infty} \frac{(-1)^m \Gamma(\alpha+1)}{m!\,\Gamma(\alpha-m+1)} - \sum_{m=0}^{\infty} \frac{(-1)^m \Gamma(\alpha+1)}{m!\,\Gamma(\alpha-m+1)} \sum_{i=1}^{S} \big(1-(1-p_i)^m\big) \\
&= \sum_{m=1}^{\infty} \frac{(-1)^{m+1} \Gamma(\alpha+1)}{m!\,\Gamma(\alpha-m+1)} S_m \\
&= \sum_{m=1}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(1-\alpha)} S_m
\end{aligned}$$

where $\Gamma$ denotes the gamma function. Hence,

$$D_\alpha = \bigg( \sum_{m=1}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(1-\alpha)} S_m \bigg)^{\frac{1}{1-\alpha}}. \tag{S8}$$

We express the link with the rarefaction curve in terms of the Tsallis entropies $T_\alpha$ (Tsallis, 1988),

$$T_\alpha = \frac{1}{1-\alpha} \Big( \sum_{i=1}^{S} p_i^\alpha - 1 \Big),$$

which is closely related to the Hill diversities $D_\alpha$,

$$D_\alpha = \big(1+(1-\alpha)T_\alpha\big)^{\frac{1}{1-\alpha}}. \tag{S9}$$

Equation (S8) becomes

$$\begin{aligned}
T_\alpha &= \frac{1}{1-\alpha} \bigg( \alpha - 1 + \sum_{m=2}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(1-\alpha)} S_m \bigg) \\
&= -1 + \sum_{m=2}^{\infty} \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(2-\alpha)} S_m.
\end{aligned}$$

We study the behavior of the coefficients $c_m$ in this infinite sum,

$$c_m = \frac{\alpha\,\Gamma(m-\alpha)}{m!\,\Gamma(2-\alpha)}.$$

For $\alpha \in (0,2)$ all coefficients $c_m$ are positive, and

$$c_m \sim m^{-(\alpha+1)} \qquad \text{as} \quad m \to \infty. \tag{S10}$$

This shows that different Tsallis entropies $T_\alpha$ depend on different parts of the rarefaction curve $S_m$. For $\alpha$ close to 2, the Tsallis entropy $T_\alpha$ is mainly determined by the rarefaction

curve for small $m$. For decreasing $\alpha$, the contribution of the rarefaction curve for large $m$ increases. For the limit cases $\alpha \to 0$ and $\alpha \to 2$ the constant of proportionality in (S10) vanishes. For $\alpha = 2$ we have $T_2 = 1 - C = S_2 - 1$: the only contribution of the rarefaction curve is at $m = 2$. For $\alpha = 0$ we have $T_0 = S - 1 = S_\infty - 1$: the contribution of the rarefaction curve is entirely shifted to $m \to \infty$. This analysis also holds for the Hill diversities $D_\alpha$ because $D_\alpha$ is an increasing function of $T_\alpha$, see (S9).

As an illustration, we apply (S8) to a community with a power-law tail. That is, we consider an artificial community consisting of an infinite number of species, for which the species are arranged in decreasing order of abundance, and for which

$$p_i \sim i^{-z} \qquad \text{as} \quad i \to \infty.$$

The abundances should be summable, so we have to impose that $z > 1$. The tail of the abundance distribution determines the asymptotic behavior of the rarefaction curve,

$$S_m \sim m^{1/z} \qquad \text{as} \quad m \to \infty.$$

From (S8) and (S10) it follows that the diversity $D_\alpha$ is finite for $\alpha > \frac{1}{z}$, and diverges for $\alpha \leq \frac{1}{z}$. This can be checked directly from Definition (S5).

# Text S4

## Estimating species abundances from sample data

The Good-Turing estimators (Good, 1953) are a well-known family of frequency estimators. Here we present a compact derivation, given in Nádas (1985), which demonstrates that the Good-Turing estimators are non-parametric, that is, free of assumptions about the abundance distribution.

Let $\Theta$ be a random variable taking values between 0 and 1, with a distribution function $G(\theta)$ about which nothing is known. Suppose that $R$ is another random variable whose conditional distribution $p_M(r|\theta)$, when $\Theta$ has the value $\theta$, is binomial with parameters $M$ and $\theta$,

$$p_M(r|\theta) = \binom{M}{r} \theta^r (1-\theta)^{M-r}. \tag{S11}$$

Then we have the identity

$$\theta \; p_M(r|\theta) = \frac{r+1}{M+1} \; p_{M+1}(r+1|\theta) \tag{S12}$$

Suppose now that we wish to estimate the value of $\theta$ given that $R$ is observed to take the value $r$. Taking a Bayesian approach with prior distribution $G$, the posterior mean for $\theta$ is

$$E[\theta|R=r] = \frac{r+1}{M+1} \, \frac{p_{M+1}(r+1)}{p_M(r)} \tag{S13}$$

where $p_M$ is the unconditional probability mass function of $R$ (that is, integrated out over $G$). This derivation is non-parametric in that $G$ is not only unknown, but no assumptions are made about $G$: the probability mass function $p_M$ must therefore be estimated directly from the sample data, so that we are in fact performing empirical Bayes estimation.

In the context of diversity estimation, we regard $G$ as the community abundance distribution, $\theta$ as the species abundance to be estimated and $r$ as the number of times that this species occurs in the sample. We use the maximum likelihood estimates for $p_M(r)$ and $p_{M+1}(r+1)$ given by $F_r/M$ and $F_{r+1}/(M+1)$, respectively. Plugging the estimates into (S13) and assuming that $M \gg 1$, we get the estimated community abundance $\widehat{\theta}_r$ of a species observed $r$ times in the sample,

$$\widehat{\theta}_r = \frac{r+1}{M} \, \frac{F_{r+1}}{F_r}, \tag{S14}$$

which are the Good-Turing frequency estimators.

As a corollary of (S14) we get the estimator for the total abundance of the observed species,

$$\sum_{r\geq 1} F_r \, \widehat{\theta}_r = \frac{1}{M} \sum_{r\geq 1} (r+1) F_{r+1} = \frac{M - F_1}{M},$$

so that the total abundance $p_{\text{unobs}}$ of the unobserved species is estimated as

$$\widehat{p}_{\text{unobs}} = \frac{F_1}{M}. \tag{S15}$$

In words, the total relative abundance of unobserved species in the community is estimated as the total relative abundance of singletons in the sample.

# Text S5

## Estimating Hill diversities from sample data

We construct estimators for the Hill diversity $D_\alpha$ based on a sample of size $M$ taken from the community. Our strategy consists in first estimating the rarefaction curve $S_m$ and then using the link (S8) between $D_\alpha$ and $S_m$.

The estimation of the rarefaction curve decomposes into two parts. For the part $m \leq M$ the rarefaction curve can be estimated unbiasedly using the estimator (S2). For the part $m > M$ the sample data have to be extrapolated, and no unbiased estimator exists. We denote the relative abundances of the unobserved species by $q_1, q_2, \ldots$ (there are $S - S_{\text{obs}}$ unobserved species). If we knew the abundances $q_i$, then we could compute the rarefaction curve using the formula,

$$\widehat{S}_m = S_{\text{obs}} + \sum_{i \geq 1} \Big(1 - \big(1 - q_i\big)^{m-M}\Big) \qquad\qquad m > M. \tag{S16}$$

As we have argued in the main text, the sample data contain little information about the abundances $q_i$ of unobserved species. However, the Good-Turing estimator (S15) for the total abundance $p_{\text{unobs}} = \sum_{i \geq 1} q_i$ of the unobserved species is available. It follows from (S16) that the estimation of the rarefaction curve $S_m$ for $m > M$ reduces to distributing the estimated abundance $\widehat{p}_{\text{unobs}}$ over the individual unobserved species.

We work out two scenarios, see Figure 3 of the main text. In the first scenario we distribute $\widehat{p}_{\text{unobs}}$ so as to obtain the lowest possible value of the diversity $D_\alpha$ *consistent with the sample data*. By this we mean that $\widehat{p}_{\text{unobs}}$ must be distributed in a manner which remains consistent with the estimates $\widehat{\theta}_r$. The lowest diversity occurs when all unobserved species have the same abundance, $q_1 = q_2 = \ldots = q^-$, and this abundance is as high as possible. However, as noted in Good (1953), the frequency estimates $\widehat{\theta}_r$ must increase as $r$ increases: this implies an upper bound for $q^-$, namely $\widehat{\theta}_1$ (which is the estimated community abundance of any species observed exactly once in the sample). We therefore take $q^- = \widehat{\theta}_1 = \frac{2F_2}{MF_1}$ so that, from (S15), there are $\frac{F_1^2}{2F_2}$ unobserved species. Hence, the estimated rarefaction curve (S16) becomes

$$\widehat{S}_m^- = S_{\text{obs}} + \frac{F_1^2}{2F_2}\left(1 - \left(1 - \frac{2F_2}{MF_1}\right)^{m-M}\right) \qquad\qquad m > M, \tag{S17}$$

where the superscript in $\widehat{S}_m^-$ indicates the low-diversity scenario.

In the second scenario we distribute $\widehat{p}_{\text{unobs}}$ so as to obtain the highest possible value of the diversity $D_\alpha$. The highest diversity is obtained when all unobserved species have the same abundance, $q_1 = q_2 = \ldots = q^+$, and this abundance is as small as possible. The smallest abundance a species can have in a community of size $N$ is equal to $\frac{1}{N}$, corresponding to a species represented by a single individual. We therefore take $q^+ = \frac{1}{N}$ so that, from (S15), there are $\frac{NF_1}{M}$ unobserved species. Hence, the estimated rarefaction curve (S16) becomes

$$\widehat{S}_m^+ = S_{\text{obs}} + \frac{NF_1}{M}\left(1 - \left(1 - \frac{1}{N}\right)^{m-M}\right) \qquad\qquad m > M, \tag{S18}$$

where the superscript in $\widehat{S}_m^+$ indicates the high-diversity scenario. Note that the upper estimator (S18) depends on the community size $N$, in contrast to the estimator (S17).

To summarize, we have obtained two estimators for the Hill diversity $D_\alpha$, a lower estimate $\widehat{D}_\alpha^-$ and an upper estimate $\widehat{D}_\alpha^+$. They can be computed as follows:

**Lower estimate** First, compute the lower estimate of the rarefaction curve. From (S2) and (S17),

$$\widehat{S}_m^- = \begin{cases} \sum_{k\geq 1} F_k \Big(1 - \frac{\binom{M-k}{m}}{\binom{M}{m}}\Big) & \text{if } m = 1, 2, \ldots, M \\ S_{\text{obs}} + \frac{F_1^2}{2F_2}\Big(1 - \big(1 - \frac{2F_2}{MF_1}\big)^{m-M}\Big) & \text{if } m = M+1, M+2, \ldots \end{cases} \tag{S19}$$

Then, substitute this result into (S8) to estimate the Hill diversity,

$$\widehat{D}_\alpha^- = \Big( \sum_{m=1}^{\infty} \frac{\alpha\, \Gamma(m-\alpha)}{m!\, \Gamma(1-\alpha)}\, \widehat{S}_m^- \Big)^{\frac{1}{1-\alpha}}. \tag{S20}$$

**Upper estimate** First, compute the upper estimate of the rarefaction curve. From (S2) and (S18),

$$\widehat{S}_m^+ = \begin{cases} \sum_{k\geq 1} F_k \Big(1 - \frac{\binom{M-k}{m}}{\binom{M}{m}}\Big) & \text{if } m = 1, 2, \ldots, M \\ S_{\text{obs}} + \frac{NF_1}{M}\Big(1 - \big(1 - \frac{1}{N}\big)^{m-M}\Big) & \text{if } M+1, M+2, \ldots \end{cases} \tag{S21}$$

Then, substitute this result into (S8) to estimate the Hill diveristy,

$$\widehat{D}_\alpha^+ = \Big( \sum_{m=1}^{\infty} \frac{\alpha\, \Gamma(m-\alpha)}{m!\, \Gamma(1-\alpha)}\, \widehat{S}_m^+ \Big)^{\frac{1}{1-\alpha}}. \tag{S22}$$

The Matlab code to compute the Hill diversity estimates $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ is part of the Supplementary Information.

We discuss three properties of the estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ that follow directly from their definitions. First, the lower estimate $\widehat{D}_\alpha^-$ generalizes Chao's estimator for species richness,

$$\widehat{D}_0^- = \widehat{S}_\infty^- = S_{\text{obs}} + \frac{F_1^2}{2F_2}.$$

Note that the lower estimate, like Chao's estimator, only gives meaningful results if the number of species observed once or twice in the sample is sufficiently large, and at least $F_2 > 0$. These conditions are typically satisfied in practice, especially for highly diverse communities.

Second, the upper estimate $\widehat{D}_\alpha^+$ depends on community size $N$, which is typically several orders of magnitude larger than sample size $M$. It is therefore instructive to consider the limit $N \to \infty$. A computation analogous to the one in Text S2 shows that the upper estimate $\widehat{D}_\alpha^+$ diverges as $N^{1-\alpha}$ for $\alpha < 1$, and as $\log N$ for $\alpha = 1$. Hence, we expect large values of the upper estimate (and therefore large estimation uncertainty) for $\alpha < 1$, especially for $\alpha$ close to zero (that is, close to species richness).

Third, the estimators $\widehat{D}_\alpha^-$ and $\widehat{D}_\alpha^+$ coincide for the Simpson diversity. The Simpson diversity $D_2$ is the only Hill diversity $D_\alpha$ that does not depend on the extrapolation of the rarefaction curve. It is a function of the rarefaction curve at $m = 2$: $D_2 = \frac{1}{2-S_2}$. Because the initial part of the estimated rarefaction curve is the same for the lower and upper estimate, the Simpson diversity estimates are equal, $\widehat{D}_2^- = \widehat{D}_2^+$. The Simpson diversity is not sensitive to the extrapolation of the rarefaction curve, and therefore easy to estimate.

# Supplementary Tables

## Table S1

Table S1: Description of communities used in Figure 2. Communities C1, C2 and C3 have a power-law abundance distribution, with parameters $S$, the number of species in the community, and $z$, the exponent of the power-law. The Hill diversity of order $\alpha = 0$ is equal to the number of species, $D_0 = S$; the Hill diversity of order $\alpha = 1$ is the Shannon diversity; the Hill diversity of order $\alpha = 2$ is the Simpson diversity. For a sample of size $2\ 10^4$, the number of observed species is denoted by $S_{\text{obs}}$ and Chao's estimator for species richness is denoted by $\widehat{S}_{\text{Chao}}$.

| | $S$ | $z$ | $D_1$ | $D_2$ | $S_{\text{obs}}$ | $\widehat{S}_{\text{Chao}}$ |
|---|---|---|---|---|---|---|
| community C1 | $5\ 10^4$ | 1.1 | 640 | 35 | $4.8\ 10^3$ | $1.5\ 10^4$ |
| community C2 | $2\ 10^5$ | 1.3 | 100 | 11 | $2.4\ 10^3$ | $8.3\ 10^3$ |
| community C3 | $10^6$ | 1.6 | 15 | 4.5 | 690 | $1.8\ 10^3$ |

# Table S2

Table S2: Data for empirically-sampled microbial communities. We report the sample size $M$, the number of species observed in the sample $S_{\text{obs}}$, the number of singleton species $F_1$, that is, the number of species that have been sampled only once, the estimated relative abundance of the unobserved species $\widehat{p}_{\text{unobs}}$, and the Chao estimate $\widehat{S}_{\text{Chao}}$ for the number of species in the community. The data sets are taken from Quince *et al.* (2008): a seawater bacterial sample from the upper ocean (Rusch *et al.*, 2007), soil bacterial samples at four locations: Brazil, Florida, Illinois and Canada (Roesch *et al.*, 2007), and seawater samples from deep-sea vents at two locations: FS312 and FS396, separated into bacteria and archaea (Huber *et al.*, 2007).

| | $M$ | $S_{\text{obs}}$ | $F_1$ | $\widehat{p}_{\text{unobs}}$ | $\widehat{S}_{\text{Chao}}$ |
|---|---|---|---|---|---|
| upper ocean | 7068 | 811 | 311 | 0.044 | 1038 |
| soil, Brazil | 26079 | 2880 | 1176 | 0.045 | 4604 |
| soil, Florida | 28150 | 3440 | 1541 | 0.055 | 5643 |
| soil, Illinois | 31621 | 3357 | 1466 | 0.046 | 5745 |
| soil, Canada | 52773 | 5515 | 2634 | 0.050 | 10394 |
| FS312, bacteria | 442062 | 12183 | 5339 | 0.012 | 19568 |
| FS312, archaea | 200199 | 1594 | 460 | 0.002 | 2175 |
| FS396, bacteria | 247826 | 5843 | 2825 | 0.011 | 10570 |
| FS396, archaea | 16428 | 418 | 158 | 0.010 | 630 |

# Supplementary Figures

## Figure S1

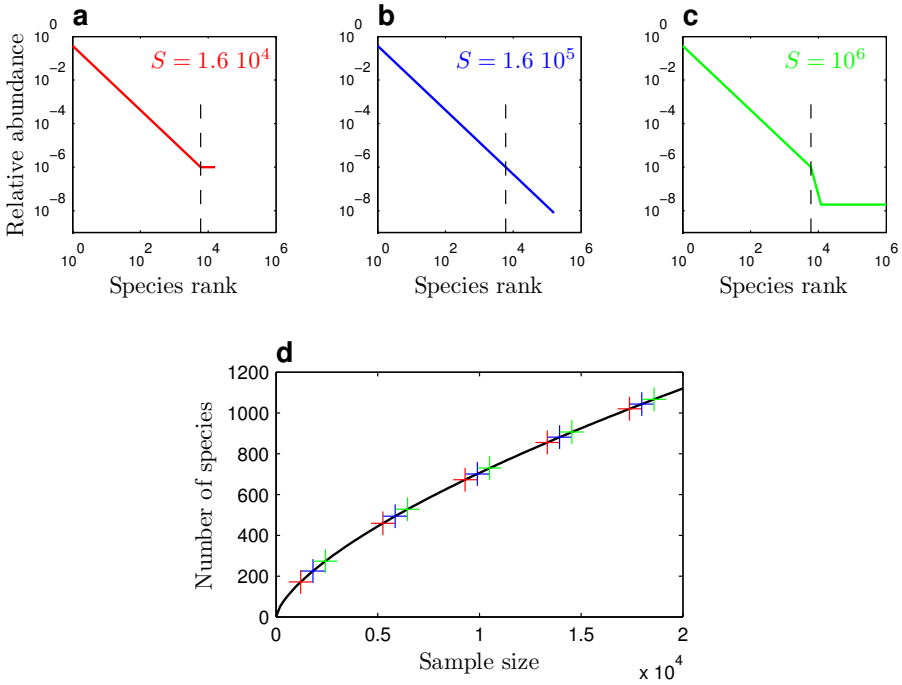


Figure S1: Sample data are insensitive to rare species tail of community. We generated three community abundance distributions, shown in red, blue and green (panels a–c). The three communities have the same abundance distribution for species with relative abundance above $10^{-6}$ (the part of the rank-abundance curve to the left of the dashed black line). This common part consists of $6\ 10^3$ species, occupying 99% of the community abundance. The communities differ in the tail of rare species: the community in panel a has $1.6\ 10^4$ species; the community in panel b has $1.6\ 10^5$ species; the community in panel c has $10^6$ species. Despite the marked differences, the rarefaction curves of the three communities up to sample size $2\ 10^4$ are identical (see panel d). This observation holds generally: any set of rare species leads to the same rarefaction curve if each rare species has relative abundance below $10^{-6}$ and the total relative abundance of the set of rare species equals 0.01.

# Figure S2

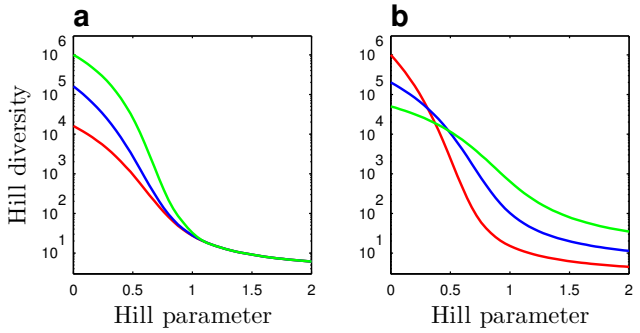


Figure S2: Hill diversity for large $\alpha$ is insensitive to rare species tail. Panel a: We computed the Hill diversity $D_\alpha$ for the three communities of Figure S1. The Hill diversities for $\alpha > 1$ almost coincide because the communities have the same set of non-rare species. The Hill diversities for $\alpha < 1$ differ because the communities have different rare species tails. Panel b: We computed the Hill diversity $D_\alpha$ for the three communities of Figure 2. The curves of Hill diversities intersect. For small $\alpha$, the most species-rich community (C3, green) has the largest Hill diversity, and the most species-poor community (C1, red) has the smallest Hill diversity. For larger $\alpha$, the most even community (C1, red) has the largest Hill diversity, and the most uneven community (C3, green) has the smallest Hill diversity.

# Figure S3

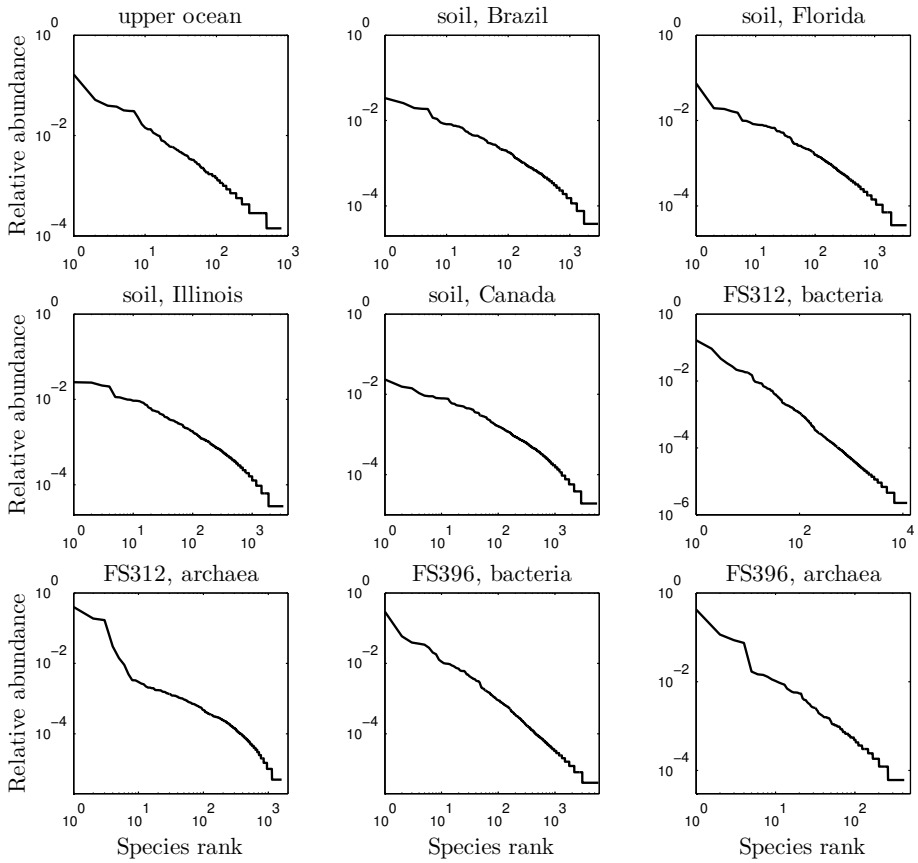


Figure S3: Rank-abundances curve of empirical microbial community samples. Relative abundance in the sample is plotted against species rank in the sample. We used the same data sets as Quince *et al.* (2008): a seawater bacterial sample from the upper ocean (Rusch *et al.*, 2007), soil bacterial samples at four locations: Brazil, Florida, Illinois and Canada (Roesch *et al.*, 2007), and seawater samples from deep-sea vents at two locations: FS312 and FS396, separated into bacteria and archaea (Huber *et al.*, 2007).

# Figure S4

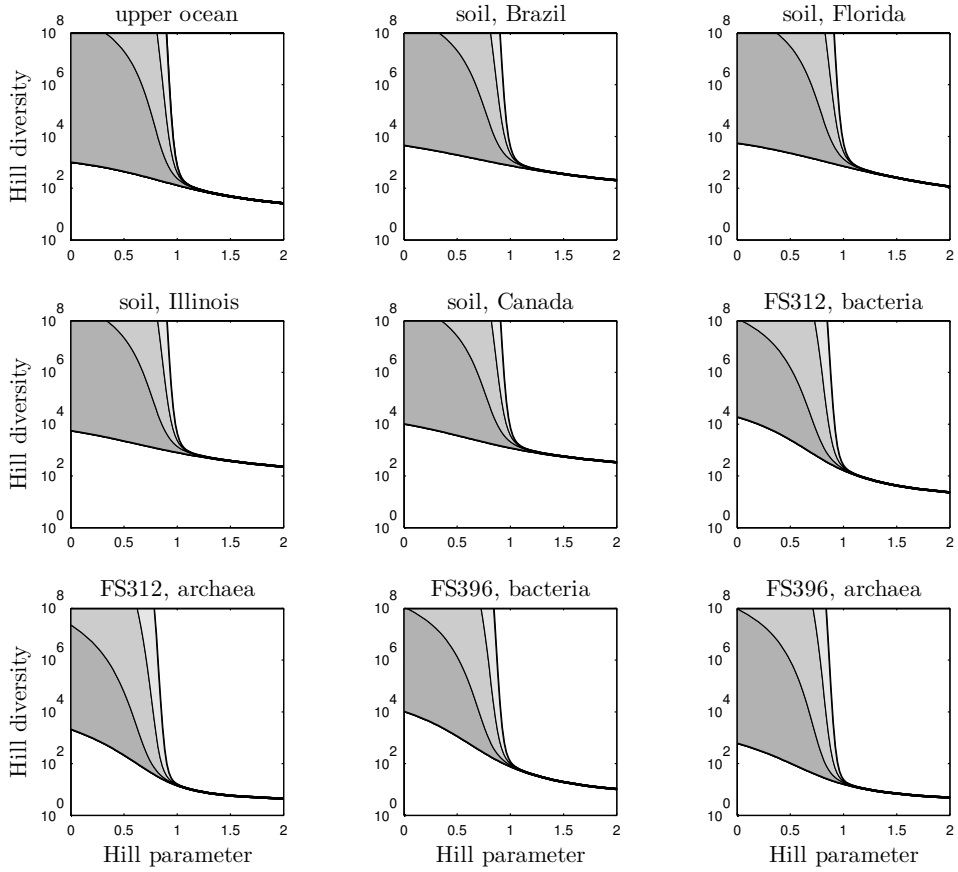


Figure S4: Community-size dependence of Hill diversity estimates. Same data sets as in Figure 5, but for three values of community size $N$. The lower estimate is independent of $N$; the upper estimate increases with increasing $N$ (from left to right: $N = 10^{10}$, $N = 10^{15}$, $N = 10^{20}$). We observe the same behavior as for the *in silico* generated data sets of Figure 4.

# References


Good IJ. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**: 237–264.

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Mao CX. (2007). Estimating species accumulation curves and diversity indices. *Statist Sinica* **17**: 761–774.

Nádas A. (1985). On Turing's formula for word probabilities. *IEEE Trans Acoust Speech Signal Processing* **33**: 1414–1416.

Quince C, Curtis TP, Sloan WT. (2008). The rational exploration of microbial diversity. *ISME J* **2**: 997–1006.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S *et al.* (2007). The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Tsallis C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *J Stat Phys* **52**: 479–487.